To appear in: E. Salas, C. Bowers & E. Edens (Eds.), *Applying resource management in organizations: A guide for training professionals*. Mahwah, NJ: Lawrence Erlbaum Assoc.

Evaluating Resource Management Training

Robert W. Holt, Deborah A. Boehm-Davis, J. Matthew Beaubien
George Mason University

Resource management is a critical component of job performance in a number of domains. Although a fair amount of research has been devoted to the development of resource management training programs (Helmreich & Foushee, 1993; Wiener, Kanki, & Helmreich 1993), much less effort has been devoted to their evaluation.

The evaluation of a training program is important for a number of reasons, not the least of which is to ascertain whether the organization's investment pays off in terms of performance improvements (Goldstein, 1993). From a cost-benefit perspective, if performance does not improve relative to the cost of implementation, then the training program should be discontinued or modified. In many domains, however, performance changes are difficult to measure because of uncontrollable factors that exist within the larger organizational context. Therefore, it is critical to develop a list of *targeted* changes in knowledge, skills, and/or attitudes that are expected to occur after training, and to investigate these *systematically* in order to weigh the costs and benefits of training (Kraiger, Ford, & Salas, 1993).

It is also important to determine where to look for changes in performance. Kirkpatrick (1976) suggested that there are different levels of analysis at which training effectiveness can be manifest: the individual, the team or crew, and the organization. A majority of the resource management literature focuses exclusively on the transfer of trained material at the individual or team level. This is quite reasonable, as individual/team *behaviors* are most directly under the control of the trainees. However, aggregate *performance* data, for example at the department or organizational level, are also important to the organization. Unfortunately, performance data, unlike measures of behavior, are frequently beyond the control of the individuals or team. For example, an aircrew may manage a crisis situation perfectly, yet factors beyond their control, such as faulty equipment, can nonetheless lead to a disaster. As such, performance data are subject to extensive confounds, and extreme care must be used when interpreting these results as evidence for or against the effectiveness of training (Campbell, 1990). For these reasons, the effects of resource management training should be evaluated in a systematic, step-by-step fashion.

This chapter will outline the steps needed to evaluate the effectiveness of a resource management training program and highlight both practical and theoretical issues that arise during this process. We will first cover general requirements for defining, implementing, and evaluating resource management training. Then we will illustrate these principles by applying them to Crew Resource Management (CRM) in the aviation domain. This chapter will focus on the application of statistical techniques and research methodology in this domain. For a more comprehensive treatment of these subjects, interested readers should consult Campbell and Stanley (1963), Cook and Campbell (1979), Howell (1997), and Pedhazur and Pedhazur Schmelkin (1991).

DEVELOPING A RESOURCE MANAGEMENT EVALUATION PLAN

Principles of Evaluation

Several different approaches are available for evaluating the effectiveness of a resource management training program (Joint Committee, 1994; Guttentag & Struening, 1975). Certain key principles underlie all these approaches. The objective of all these approaches is to find out (1) if resource management training makes any noticeable difference in the dependent variables, and (2) what the size of the training effect is.

At a minimum, training should make a difference that is *noticeable*. A noticeable difference has two components. First, it is a difference that statistical methods determine to be non-chance (above a background level of noise due to measurement errors, individual differences, and so forth). Second, the difference should have a practical value or utility to the organization. If it is determined that training made a noticeable difference, the size of the training effect should be estimated so that cost-benefit analyses can be performed. If multiple training programs have been developed, the data can be used to assess the *relative* effectiveness of the different methods.

When evaluating training, it is critical to collect measures of performance at the appropriate time. If performance is evaluated “too soon,” before training has “sunk in,” a training effect may not be observed (Kraiger et al., 1993). Similarly, evaluating performance after too long an interval may clutter the picture with uncontrolled intervening events that affect performance and obscure the effects of training. Thus, the right time interval must be chosen to accurately evaluate training effects. If a valid theory of performance is available for the training domain, the time interval can be based on this theory. Alternatively, if the right time interval is not known, evaluation should be repeated over a reasonable period of time to check for immediate or delayed training effects.

In evaluating whether a training program made a difference, it is also important to remember that any measured effect can have multiple causes. Although training is one such cause, systematic evaluation should rule out plausible alternatives so that we can be sure that the training *per se* is making the difference. The alternative factors that could create an observed effect or difference are called “confounds” (Cook & Campbell, 1979). The evaluation of resource management training must take potential confounds into consideration, primarily in order to rule them out as alternative explanations of the observed effects.

Selecting an Evaluation Design

All evaluations of resource management training programs rely on comparisons. The simplest form of evaluation involves comparing trained and untrained groups on the same set of criteria. Alternatively, groups with different degrees or methods of training can be compared. Still another form of comparison is to compare performance after training to performance before training (baseline performance). Although all systematic evaluation relies on some form of comparison, different approaches differ on the type of comparison emphasized and on the amount of control over confounds afforded. The goal is to develop the fairest and least confounded comparison of the effects of different levels of training (Campbell & Stanley, 1963; Cook & Campbell, 1976).

Different approaches to comparison vary in the degree of intervention and control of the situation that is required for the evaluation. Evaluation approaches range along a continuum from extremely controlled studies modeled on the experimental method, with people randomly assigned to separate trained and untrained groups, to relatively uncontrolled studies, such as field studies, in which the training is done naturally and the effects measured in the natural environment. There are costs and benefits associated with each approach. In general, more controlled evaluations offer more precise information about the effects of training because they isolate changes due to the training itself rather than other extraneous factors (confounds). However, such evaluations are more difficult to set up and execute.

Experimental Designs

A traditional experimental design requires the ability to randomly assign persons to trained and untrained groups (or different levels/types of training). The trained group is compared to the untrained group on each possible effect. This is the most precise evaluation of training effectiveness, but probably also the least practical, as most organizations will usually want to apply human resources training to all job incumbents. One possible variation of the traditional experiment that may be more feasible is a “waiting list” control group. In this variation, all people ultimately receive the training, but the people designated to get the training first vs. last are randomly determined. In the window of time where the first group(s) are trained and the last group(s) are not, the effects of training can be measured on what are essentially randomly-assigned trained and untrained groups.

Quasi-experimental Designs

If naturally occurring groups are available but cannot be randomly assigned, a quasi-experiment can be constructed in which one group is trained and the other group is not. Both groups are evaluated for the effects of resource management training in the same manner as in a traditional experiment. The major disadvantage of this design, however, is that the groups may not be equivalent on other relevant variables such as ability, experience, etc. In aviation, the naturally occurring groups are fleets, and fleets may be different in the average age and experience of pilots. The possibility exists that some confounds unique to the trained group interact with the training to produce any measured effects. This makes it essential to measure possible confounds and attempt to assess their effects on the evaluation criteria.

Pre-Post evaluation

If everyone must receive human resource training at the same time, evaluation can still be based on a pre/post or time-series design. A pre/post training design examines differences before and after training for a single group. Resource management training should increase good outcomes such as efficiency, productivity, etc. while decreasing mistakes, errors, or other poor outcomes. This is one of the easiest methods of evaluation, and at the very minimum, a pre/post design should be used to evaluate the effects of training. There are some notes of caution in using this design. If performance is measured pre- and post-training, the pre-training measurement should be taken early enough to be unaffected by the knowledge of or anticipation of training, but not so early that the baseline performance could change a great deal prior to training. Unfortunately, this evaluation method is also one of the weakest since it is subject to many confounds such as contextual events, natural development, or change in the trainees. If these confounds occur between the pre-training and post-training measurements, they can artificially cause the observed changes. If these confounds are likely, a stronger form of evaluation such as a time series should be used if possible.

Time-series evaluation

A time-series design extends the time where performance is measured before and after training. Extending these intervals of measurement gives the advantage of being able to rule out potential confounds such as a general increase in the effects due to natural development over time. However, it does this at the cost of additional measurements. When making multiple measurements, the effect of the measurement process itself must be considered. If supervisors are simply rating subordinates based on naturally observed performance, there may be no reactivity in measurement on the part of the subordinate (although effects of making multiple assessments on the part of the supervisor should still be considered). However, if subordinates are put in a testing or evaluation scenario for each measurement, then practice effects, learning of test-relevant knowledge and skills, or decreases of motivation to perform well may occur. Any situation in which the subordinate is strongly aware of the testing and evaluation process is open to these types of confounds.

DEVELOPING MEASURES OF RESOURCE MANAGEMENT PERFORMANCE

Once the research design is selected, measures of performance that address the questions of interest must be developed. Accurate performance assessment requires several critical steps: defining the construct, developing appropriate measurement methods, and objectively confirming the sensitivity, reliability, and validity of these measures. These steps are interdependent.

To accurately assess resource management, it must first be defined. Without a specific operational definition, appropriate assessments of resource management cannot be developed. If the construct is multi-dimensional, then multiple measures need to be developed. Once developed, these measures must be evaluated for acceptable levels of statistical sensitivity, reliability, and validity. After the quality of these measures has been established, they may be confidently used to obtain a full and accurate evaluation of the current state of resource management, and the effectiveness of resource management training.

Defining Resource Management

Resource management is potentially difficult to define and measure because it is complex, multi-dimensional, and process-oriented (see Lauber, 1984 for more information). Given this complexity, it may be necessary to create several specific operational definitions that focus on distinct *subsets* of resource management dimensions and processes.

An operational definition of a construct is a specific, focused definition that is used for a specific purpose, such as evaluation. Any operational definition must be complete and specific enough to clearly imply appropriate measurement strategies and techniques. As a general rule, the operational definition should specify the core knowledge, skills, and behaviors needed for effective resource management in that domain. It may also include other relevant personal factors such as personality, motivation and attitudes as well as relevant situational factors.

Developing Appropriate Measures

Good resource management can affect both task (e.g., technical) and contextual (e.g. relationship-oriented) aspects of performance (Borman & Motowidlo, 1993). The performance changes can occur at the individual, team, or organizational level, and may be qualitatively different at each level (Kirkpatrick, 1976). Thus, a comprehensive evaluation process would include measures to assess both task and contextual performance. However, practical limitations generally require the evaluation process to focus on a selected subset of these possible effects. This subset should include performance improvements at the individual, team or organizational level as well as measures reflecting effective resource management processes.

The changes at the individual, team, or organization level may occur at different time frames. Kirkpatrick (1976) proposed a model which suggests that training results are manifested at multiple stages: initial reactions to the training program, changes in knowledge and behavior during the training, transfer of trained behaviors to the workplace, and changes in organizational effectiveness. According to his model, each stage is a necessary but insufficient precursor to the following stages. Despite both theoretical and empirical criticisms and caveats (Alliger & Janak, 1989; Alliger, Tannenbaum, Bennett, Traver, & Shotland, 1997; Goodman, Lerch, & Mukhopadhyay, 1994), this model provides a useful framework for considering the effect of training interventions at different organizational levels. Generally, individual effects of training may appear first, followed by team performance changes and then organizational changes. Therefore, the appropriate time to measure individual, team, and organizational effects may be quite different.

Unfortunately, measurement of resource management performance is more difficult than, say, measuring the output of an assembly-line worker. When a physical object is being produced, productivity can be indexed in terms of output quality or quantity. In contrast, the evaluation of process variables such as resource management requires evaluating the interaction of a person or team within a complex system. For example, evaluating resource management in aircrews depends on the interaction between the Captain and the First Officer as well as interaction with flight attendants, air traffic control personnel, the physical aircraft systems, and so forth (Boehm-Davis, Holt, & Seamster, this volume).

One important principle of measurement is to measure a construct with different methods. The principle of converging operations (Campbell & Fiske, 1959) is that if different measurement methods give the same result, the confidence in the result is increased. Whenever possible, alternative measures of resource management performance should be designed.

Since training may have multiple effects relevant to the organization, it is also wise to measure more than just one possible effect of training (Kraiger, Ford, & Salas, 1993). Other relevant effects of resource management training at the individual level may be attitudes toward good resource management, knowledge of resource management techniques or procedures, and the perceived effects of good resource management. Relevant team effects may be increased team morale, improved communication, etc.

Measuring Performance

Due to the complexity of resource management performance, the evaluation method of choice is often a rating by another person of the quality of resource management behavior. This evaluation should be guided by appropriate tools and materials that help the evaluator make an accurate assessment. For example, carefully designed evaluation scales and evaluation worksheets, developed according to human factors principles, have the potential to reduce the rater's cognitive workload. This may simplify the evaluation process and give more reliable results (see Boehm-Davis, Holt, & Seamster, this volume). Other materials required for evaluation will depend on the evaluation context.

The context for evaluation can be either normal daily performance on the job or a special evaluation context. One common method is to have evaluators make an overall assessment of typical performance once, at the end of the year. These evaluators can be supervisors, peers, or subordinates (i.e., 360-degree evaluation). This annual evaluation has the advantage of reflecting the person's resource management in diverse job-related situations over an extended time period. In such cases, materials beyond an evaluation scale and worksheet are not typically used. Nevertheless, there are disadvantages when using this evaluation technique. These include: incomplete or distorted recall for relevant events, recency bias, memory priming caused by the phrasing of evaluation questions, and the influence of established knowledge about the person being evaluated (DeNisi, Cafferty & Meglino, 1984).

Other evaluation problems depend on the number of persons evaluated. If the evaluator rates only one person, the evaluation may be poor because the evaluator has little practice and no good judgment anchors due to having no knowledge of the range of possible performance. If the evaluator assesses multiple persons, contrast or carry-over effects from one person to the next can influence the ratings.

Special evaluation contexts can be designed to avoid or minimize these errors, but may have the disadvantage that performance in the special context is at a maximal rather than a typical level, and thus would not generalize to the job (Dubois, Sackett, Zedeck, & Fogli, 1993; Sackett, Zedeck, & Fogli, 1988). However, designing "work sample" evaluations with normal job content and processes can maximize generalization. In the aviation domain, the "work sample" of a normal flight is combined with realistic simulations of normal working conditions to increase generalization and obtain more typical levels of resource management behaviors. Special evaluation requires preparation of extra materials such as the work sample itself and guides/scripts for the required behavior of the evaluator during the assessment. Furthermore, evaluators must be appropriately trained in the administration and use of these materials as well as making the critical assessments of resource management (Prince, Oser, Salas, & Woodruff, 1993).

Measuring Knowledge

Another option for evaluating resource management is to evaluate a component that contributes to performance, such as the information that individuals have assimilated into their “tool kit” as a result of the training program. Two types of knowledge have been identified that might result from a training intervention: declarative and procedural knowledge.

Declarative knowledge refers to the static information about a domain that is represented in memory. It can be thought of as the definitions for constructs in the domain, and rules for when this knowledge can be (or should be) applied. Procedural knowledge, on the other hand, typically refers to rules for the *execution* of specific behaviors (Anderson, 1985). Although procedural knowledge is based in part on declarative knowledge, it is considered to be a “higher order” form of knowledge, because it involves the integration of multiple sources of information and the automation of specific behaviors. For example, before one can turn an aircraft by coordinating aileron movements with appropriate rudder movements, one must have the appropriate foundation of declarative knowledge about adverse yaw caused by moving the ailerons. Because procedural and declarative knowledge are manifest in different forms, they must be assessed differently. Typically, declarative knowledge is assessed via paper-and-pencil measures, while procedural knowledge is assessed with some form of work sample test or measures of underlying cognitive structure (Kraiger, Ford, & Salas, 1993).

Effectiveness Criteria

Another issue to consider when developing appropriate measures is whether the goal is to change the mean (average) level of performance or to change the distribution (variability) of performance. Traditionally, training is evaluated in terms of mean differences. For example, the mean pre-training performance ratings for crews are often compared to mean post-training performance ratings. Likewise, the mean performance of trained crews is often compared to that of untrained crews. However, some researchers (Alliger & Katzman, 1997) argue that certain training interventions can influence both the mean and/or variability of performance data. For example, group consensus training or instructor calibration training is often used to decrease the random variability in people's response patterns with no change in the mean. Conversely, training may attempt to increase the variance. For example, training in creativity may seek to increase the variability of ideas generated by a group. Therefore, it is essential that researchers avoid the temptation to assess training performance solely in terms of mean change; doing so can ignore potentially valuable information regarding the training program's effectiveness. Specifying which types of changes should occur can be guided by an overall theory of resource management in the particular domain.

This theory of performance should also be used to develop a systematic measurement plan (Kraiger, Ford, & Salas, 1993). The plan should specify which types of performance should be expected at various times after the training, the level of such performance, and the appropriate measurement strategy.

Ensuring the Quality of Measurement

The third step in assessment is to objectively confirm the quality of these assessment measures. Since evaluations are performed by individuals, the quality of the evaluations is decreased by inaccuracy, subjectivity, or personal biases on the part of the evaluator. Objectively confirming good quality measurement involves three basic facets. A good measure of resource management must be *sensitive* enough to discriminate good from poor resource management, *reliable* enough to consistently provide the same estimate of resource management, and *valid* enough to ensure that the measure involves only resource management rather than other extraneous factors. We will cover each facet of good measurement in turn.

Measurement Sensitivity

Sensitivity refers to the extent to which a measure can detect changes in the construct being assessed. Specifically, a sensitive measure of resource management should show higher scores when resource management is good and lower scores when resource management is bad. Extreme examples of very good or very bad performance are usually quite easy to detect, but sensitivity must also be established for distinguishing more subtle differences in resource management behaviors, for example differences resulting in *marginally* safe vs. unsafe performance.

Sensitivity is influenced by the granularity of the measurement instrument. The evaluation scale must be sufficiently fine-grained to capture important differences in the quality of resource management that is observed, yet still be accurately used by the evaluator. For example, a simple dichotomous “good or bad” scale might be accurately used by evaluators, but would not be sensitive to different degrees or good or bad resource management. Conversely, a 100-point scale might be extremely fine-grained, but evaluators may not be able to use it accurately. A good compromise for measurements based on human evaluations is often a four or five-point scale with meaningful definitions assigned to each scale point (Likert, 1936).

To objectively index the sensitivity of measurement, it is necessary to compare the judgements made by evaluators to established levels of resource management. One method for indexing the sensitivity of evaluation is to have evaluators rate “test” cases and determine their rating levels for cases with different levels of resource management proficiency (as established by subject-matter experts). For example, average evaluator ratings for “good” test cases ought to be higher than ratings for “average” test cases, which in turn should be higher than ratings for “poor” test cases. Holt, Johnson, & Goldsmith (1996) have used Hays’ (1988) omega-squared index for strength of effect as the objective index of sensitivity. This has a range from 0 (no discrimination of different levels) to 1 (perfect discrimination of different levels).

Measurement Reliability

Informally, reliability can be thought of as the consistency or stability of measurement. Formally, reliability is defined as the lack of random error in the measurement instrument (Nunnally, 1967). Different traditional methods of estimating reliability have been developed, and we will cover two: test-retest reliability and internal consistency reliability (see Nunally, 1967 or Pedhazur & Pedhazur Schmelkin, 1991 for more information). Each traditional method makes different assumptions about the main source of error in measurement, and each method has notable disadvantages when used for estimating the reliability of human evaluations.

Test-retest reliability is used to assess the temporal stability of a measurement over time. One method of assessing this form of reliability consists of having evaluators assess the same set of performances at two different times and correlating these two sets of evaluations. The calculation uses the Pearson product-moment correlation and results in an index *r* that reflects reliability. In this case a value of *r* near 1.0 indicates near-perfect test-retest reliability whereas values near 0 indicate a lack of test-retest reliability.

However, test-retest reliability assumes that the only important source of random error is spontaneous changes over time. Unfortunately, systematic evaluator differences are common in evaluating resource management in the aviation domain (Williams, Holt, & Boehm-Davis, 1997). To the extent these differences are stable over time, the test-retest reliability is inflated. Therefore, although simple to execute, the test-retest reliability method only covers one potential source of error and may be positively biased.

Internal consistency reliability refers to the internal coherence or intercorrelations of a set of items which are all measuring the same thing (Nunnally, 1967). For evaluation of resource management, this type of reliability requires a set of multiple items all reflecting resource management. If resource management has distinct components, each distinct component must have its own set of multiple items. The intercorrelations among items in a set are summarized into a Coefficient Alpha index which ranges from 0 (no internal consistency reliability) to 1 (perfect internal consistency reliability). Several factors influence coefficient alpha, such as the number of items included in the scale (Cortina, 1996). Furthermore, this form of reliability also ignores some forms of *systematic* judgment errors made by evaluators (e.g., halo rating errors). To the extent that these systematic errors occur across items, the internal consistency is inflated.

Thus, when used in isolation, both test-retest and internal consistency reliability estimates can provide misleading results due to the occurrence of rater errors. To check and correct such rater errors, we developed an alternative approach for training and checking evaluator reliability using multiple statistical indexes for evaluating rater performance and giving training feedback. This multi-component approach was labeled Inter-Rater Reliability training (IRR).

During the IRR process, each evaluator's ratings of the test cases are compared to the group's judgments by using four indexes, each of which provides information on one aspect of reliability. In addition, an index of the sensitivity of judgment is included. First, the overall distribution of each evaluator's ratings is compared to the group's distribution to ascertain its level of *congruency*. Low congruency suggests the evaluator gives a different mix of ratings on the scale compared to the group. Second, *systematic differences* of harsher or more lenient grading among the evaluators are identified.

Third, the inter-rater correlation is calculated to see if the raters shift in a consistent manner up and down in their ratings across evaluated items. Fourth, if the test cases have been externally scored by subject-matter experts, the raters can also be assessed on the *sensitivity* of their evaluations as discussed earlier. Congruency, systematic differences, consistency, and sensitivity results are given to individuals, and the aggregate results for all raters are reported to the group.

The group of raters is also given the level of group *agreement* on each item. This feedback is critical because every item with low agreement should be discussed until reasonable group consensus is reached. In summary, the IRR method compares each rater to the group using indexes that give the rater information about the congruency, systematic differences, consistency and sensitivity of his or her evaluations. The information from these indexes is then used to improve group agreement in ratings (Williams, Holt, & Boehm-Davis, 1997)..

Measurement validity

Validity is the extent to which a measure really measures the intended construct (Nunnally, 1967; Landy 1986). More specifically, validity is the proportion of variance in a measure that reflects real variation in the measured items. From a resource management perspective, validity refers to the amount of variability in evaluator ratings that is accurately reflecting real differences in the resource management performance of the persons being evaluated. Assessing validity requires checking measurement items, the measurement process, and the results of the measurement process.

The items used for evaluating resource management should be checked for face and content validity (Nunnally 1967). Face validity is the judgment of a group of experts that the items are plausibly measuring the desired construct. Such judgments are easy and convenient, but unfortunately they are also somewhat subjective. A more objective item analysis will often indicate that items designed by experts to measure a given construct do not, in fact, predict that construct. Face validity is, therefore, easy to establish but only weak evidence for validity.

Content validity first requires a careful specification of the domain of all possible relevant items. Content validity can then be demonstrated by showing that the evaluation items are a fair, unbiased, and representative sample of items from this larger domain. Techniques for specifying relevant content items for training programs have been developed (Lawshe, 1975). However, resource management typically requires an individual or team to interact in a complex system and the set of possible items is very large and ill-defined. Therefore, the specification of the domain of all possible relevant items for this type of complex domain may be difficult or impossible.

The validity of measurement generally is established by empirically examining the relationship to other measures that should be related to the construct. Two basic principles apply. The first principle is convergent and discriminant validity (Campbell & Fiske, 1959). In convergent validity, measures which ought to be related to a construct should converge or correlate with the proposed measure. For resource management, measures which ought to positively relate to it, such as measures of general cognitive ability, ought to positively correlate with the resource management measure. A valid measure of a construct should show the expected relationships with plausible criteria (criterion validity), and predict the expected outcomes of changing resource management (predictive validity).

In divergent validity, measures which ought to be independent or distinct from resource management should diverge or not correlate with the proposed measure. For example, if resource management can be done equally well by men and women, then gender should *not* correlate with resource management measures. Divergent validity is particularly important if potential confounds like popularity or appearance could influence a measure of resource management effectiveness.

The second principle is network validity (Pedhazur & Pedhazur Schmelkin, 1991). For network validity, the nomological network of constructs that should be theoretically associated with the construct is empirically assessed to determine if it demonstrates the expected pattern of relationships. For example, a valid measure of resource management ought to show a plausible set of relationships with antecedents, concomitants, and consequences that one would expect for resource management. If the expected network of relationships is generally found, network validity is established.

EXAMPLE EVALUATION PROGRAM DEVELOPMENT

As part of a research project, we recently worked with a regional air carrier to develop and evaluate a resource management training program for pilots. This training program focused on improved crew briefings and communication during normal operations and better problem diagnosis, situation assessment, planning and decision making during abnormal or emergency operations. This program was unique in that the resource management principles were translated into step-by-step operational procedures. Further, these procedures were formally required as part of Standard Operating Procedure (SOP) for one fleet and added to their operating manuals and handbooks for that particular aircraft type.

Selecting the Level at which Resource Management would be Evaluated

For this project, we decided to evaluate the effectiveness of the training program by measuring performance at both the individual pilot and crew levels. Clearly, the performance of the individual pilot is important. First, individual pilots must be qualified to continue to legally operate an aircraft. Second, the performance of an individual can directly affect the performance of his or her team or crew. Third, some issues such as the effects of ability on performance, were more sensibly addressed by comparing the assessed ability of individuals to their performance (Boehm-Davis, Holt, & Hansberger, 1997).

Although individual performance is important, commercial aircraft are always operated by crews. The performance of a team or crew may be quite distinct from the potential individual performances of team members (Steiner, 1972). Evidence from aviation accidents and safety reports suggests that a lack of coordination among crewmembers has been the cause of numerous problems on the flight deck (NTSB, 1994). Thus, this project also focused on crew resource management performance.

Developing the Evaluation Plan

Selecting an Evaluation Design

The airline in question was composed primarily of two fleets. The research team decided to provide the resource management training program to one of the fleets, while the other fleet continued to use existing procedures and management techniques. In this quasi-experimental design, the fleet with extra training and new procedures acted as the experimental group while the fleet with normal training and procedures acted as the control group. One focus of the evaluation design was to compare pilots and crews in the two fleets.

In order to allow for gradual learning on the part of the pilots of the new procedures and processes, we also incorporated aspects of a time-series design. We collected pilot and crew performance measures over a three-year period. During the first year, the pilots had additional resource management training but the new procedures had not been formally implemented. This was our baseline performance year. In the second year, the new procedures were formally implemented and required as Standard Operating Procedure for that fleet. Performance measured in that year would reflect the immediate impact of the resource management training and SOP changes. The third year was the final follow-up assessment that would confirm or disconfirm long-term effects of the training, including a gradual acceptance and accommodation to the new methods of cockpit interaction and coordination.

In addition, during the final third year of evaluation, three auxiliary measures of resource training were developed. These additional methods allowed a converging measurement of the effects of this training with different samples of evaluators and performance situations.

Developing Measures of Resource Management Performance

Once the evaluation design had been selected, the next steps were to develop an operational definition of resource management, develop appropriate measures given that operational definition, and to ensure the quality of the measures that were developed.

Defining Resource Management

For this project, effective crew resource management (CRM) was defined for two qualitatively distinct contexts: normal operations and abnormal/emergency situations. For normal situations, effective CRM was defined as the effective communication and coordination of crewmembers before, during, and after flying a typical flight. The operational definition of normal performance included quality of briefings and other communication, quality of workload management and avoiding overload, maintaining situation awareness of the aircraft and external traffic and weather situation, and preserving effective coordination on checklists, flows and other sequential tasks during the flight.

For abnormal/emergency situations, effective CRM was defined as effective workload management and communication while performing normal flight tasks plus problem diagnosis, situation assessment, planning, and monitoring of plan execution. The operational definition of abnormal/emergency CRM was quite extensive and included, for example, the establishment of explicit “Bottom Lines” and “Backup Plans” during the planning task plus clearly communicating these plan components to other crew members.

Developing Appropriate Measures

In carrying out this project, we realized the need to develop a variety of measures to capture both individual and team level performance. Further, we realized that the metrics would be applied by a number of different evaluators (pilot instructor/evaluators). Thus, we felt that it was also important to develop a structured method for collecting the assessments of pilot performance.

Measuring performance. A structured evaluation process was designed to achieve systematic and reliable observations and ratings of performance. The multi-year evaluations consisted of a Line Operational Evaluation (LOE) and Line Checks. The LOE evaluation, conducted during the pilots’ annual evaluation for flight certification, was a work-sample performance evaluation in which the crews performed a typical carrier flight scenario in a full-motion simulator. The evaluator followed an LOE script to consistently introduce specific problems and distracting conditions into the flight. In this way, the crew reactions to abnormal and emergency situations could be assessed in a standardized manner. The evaluation forms emphasized specific crew reactions for these events, including both technical and CRM performance items.

The basis for the evaluation forms was the specification of a set of observable behaviors. These observable behaviors were carefully identified by experts as being central to successful performance on a specific event set. These behaviors also provided a point of focus for the instructor/evaluators during the observation of the LOE and during the crew debriefing after the evaluation.

The Line Check assessed pilot and crew performance during normal flight operations. Instructor/evaluators would get on a routine flight without prior announcement and evaluate the crew on a spectrum of technical and CRM items. For this carrier, crew performance ratings, both technical and

CRM, were based on a standardized four-point scale covering the full range of possible crew performance from Unsatisfactory Performance (observed crew behavior does not meet minimal requirements) to Above Standard Performance (observed crew behavior is markedly better than the Standard Performance).

To provide converging measurement of crew performance, we designed two auxiliary performance measures. First, the Instructor/Evaluators who had evaluated pilots from both the experimental fleet and the control fleet completed a detailed performance questionnaire about the relative performance of pilots from both fleets during upgrade or transition training. Second, a separate cadre of five evaluators assessed pilots from both fleets during normal flights using a direct observation form. This cadre was completely different from the carrier's Line Check evaluators or FAA evaluators, and the assessments were strictly voluntary. These additional measures gave a converging measurement check on the hypothesized performance differences of the two fleets using different sets of evaluators and different evaluation forms from the LOE or Line Check.

To ensure a broader measurement of possible training effects besides performance, we also used a pilot survey to measure knowledge and attitudes as suggested by by Kraiger, Ford, and Salas (1993). Knowledge acquired by individual pilots from the training program was measured by a survey of all carrier pilots in the final year of the project. Knowledge was only measured post-training because the training introduced completely new procedures developed for this project which pilots could not have known about previously. Therefore, the focus of the knowledge evaluation was on the extent to which individual pilots were able to describe the new set of procedures and the appropriate context for enacting each procedure.

The focus of attitude measurement was attitudes toward CRM in general and more specifically towards the trained resource management procedures. The survey also measured how often pilots performed the new procedures and briefings, and the perceived effects of the new procedures and briefings.

Focus on mean changes. The major focus in this project was on mean differences between the two fleets. That is, we were interested in demonstrating that the crews in the trained fleet would perform at a higher level on measured crew resource management skills than crews that had not received the training. Mean differences were the focus because narrowing the range or variability of performance would not have been a useful outcome.

For attitudes, we compared the mean attitudes to a neutral baseline and compared attitudes of pilots in the two fleets. For assessing knowledge, we assessed the relative extent of relevant knowledge and tested whether the trained pilots could answer knowledge questions at an above-chance level (representing *some* knowledge). Beyond this low hurdle, we assessed the relative extent of knowledge of trained pilots for the new resource management procedures.

Ensuring the Quality of Measurement

Sensitivity

Instructor/Evaluators were presented with videotapes showing different levels of resource management behavior, derived from simulation sessions conducted by the airline. The Instructor/Evaluators (I/Es) were asked to rate the level of resource management behavior exhibited on the videotapes using a four-point scale, ranging from unsatisfactory (1) through FAA standard (2), company standard (3), and above standard (4). Each level of this scale had a unique well-anchored qualitative meaning for the raters. The segments of behavior portrayed on the videotapes were selected to represent the range of possible resource management behavior, with a focus on behaviors rated in the central portion of the scale (levels 2 and 3). Subject-matter experts established the exact level of performance for each segment. Sensitivity was indexed by analyzing the differences in each rater's evaluations for performance segments at different levels.

Reliability

Reliability was assessed on a regular basis (approximately every 6 months) using the multi-dimensional inter-rater reliability procedures developed for this project (Holt, Johnson, & Goldsmith, 1997). This process relies on a group of raters (instructor/evaluator pilots) using normative information for standardizing inter-rater reliability. All raters individually evaluated a tape of typical crew performance on the LOE, and these evaluations are statistically compared. The relative amount of congruency of judgment distributions, systematic harsh or lenient judgments, inter-rater consistency, and agreement were assessed at each session. Each rater received same-day feedback about his or her evaluation performance relative to the other raters. Each single item with agreement below a corporate standard was discussed intensively by the group to isolate and solve causes of rater variability. The focus for this part of the training was reducing the random variability of the raters on each item. Information from these discussions was also used to modify the content of the evaluation scenario, re-write evaluation items for greater clarity, formally codify explicit grading standards for certain items, and to modify or clarify carrier policies and procedures (SOP).

After this training, the performance evaluations conducted by these raters were accumulated in a database. After sufficient data were collected, the items that were designed to measure the same aspect of performance were assessed by an internal-consistency reliability metric (coefficient alpha—Cronbach, 1951). These estimates served as a final check on the reliability of the performance data.

Validity

The major focus for assessing validity was the internal structural validity of the assessment process. The evaluation data were analyzed by path analysis to verify that the process of evaluation was in fact performed in the correct manner. The process of evaluation from detailed behavioral observations to judgments of performance components to overall evaluations of performance was used to construct an anticipated structure of relationships among the performance measures. The expected structure of relationships was found, which supported measurement validity.

Analysis and Interpretation of Evaluation Results

The LOE, Line Check, and auxiliary measures were all analyzed for the hypothesized fleet performance differences. Evidence from the LOE, and the direct cockpit observations were crew-level assessments, and these results were examined for mean differences in crew performance. The evaluation of individual pilots was emphasized in the Line Check evaluations, the Instructor/Evaluator survey, and in a survey of individual pilots. Across these measures both individual and crew levels of performance could be assessed, which were the targeted levels of change for this study.

Crew performance. On the LOE, several specific items concerning crew resource management behavior were graded with exactly the same grading standards for both fleets. For most of these items, the trained and untrained fleets were significantly different in the expected direction. The conclusion was that the resource management training had the desired effects for the work-sample evaluation.

The second crew-level evaluation was direct observations of cockpit interaction on regular line flights. These observations were carried out by a separate cadre of pilots who rode in the cockpit and watched the crew under voluntary, non-jeopardy conditions. Specific briefing content and other aspects performance relevant to the training were evaluated by these observers. These direct observations of cockpit interaction showed that crews from the trained fleet were significantly superior on the majority of these items. On the remaining items, the trained fleet still had a higher mean, but the observed difference was not statistically significant.

Individual Pilot performance. The first measure of performance at the pilot level was the Instructor/Evaluator survey. This survey involved a comparison of pilots from trained and untrained fleets who were transitioning aircraft or upgrading from First Officer to Captain. Instructor/Evaluators who had experience with both sets of pilots gave comparative ratings for average individual pilot performance. These ratings indicated that pilots from the trained fleet were significantly better in communication, workload management, and planning and decision-making.

The second measure of the effects of training on individual pilots was the pilot survey. The survey included pilots trained in specific resource management procedures and pilots without this training.   Trained pilots had acquired a significant amount of knowledge about the resource management procedures, had very positive attitudes toward CRM and resource management procedures, performed the trained procedures on routine flights, and perceived that the procedures increased their effectiveness.

Convergent results for performance measurement and confirmatory results for attitudes and knowledge gives more confidence in the final evaluation of the effectiveness of this type of resource management training. Multiple evaluations at both the individual pilot level as well as the crew level help rule out various confounds or alternative explanations for the results. For example, positive effects of training were reported by Instructor/Evaluators, by an independent cadre of observer pilots, and by the pilots themselves. Each of these groups has different potential sources of bias, and the convergence of results helps reassure us that the positive effects are not the result of biased evaluators.

LESSONS LEARNED THROUGH DEVELOPING EVALUATIONS OF RESOURCE MANAGEMENT TRAINING PROGRAMS

In developing a plan to evaluate a resource management training program, we recommend following the steps outlined in this chapter. These include: selecting a level at which to measure resource management, selecting a research design through which to evaluate the selected level of resource management behavior, and defining measures that can accurately assess resource management behaviors.

Specifically, Table 1 provides an overview of the steps needed to establish and implement an evaluation of a resource management training program. Each step has a set of critical issues that should be resolved for the best possible outcome.

Table 1.
Steps for Developing an Evaluation of a Resource Management Training Program

1. Select the level at which resource management will be evaluated

2. Develop the evaluation plan
    - Select an evaluation design
    - Determine appropriate time interval for measuring change

3. Develop measures of resource management performance
    - Operationally define resource management
    - Develop appropriate measures
        - Measure knowledge, attitude, performance, etc.
        - Develop converging measures where possible
        - Decide to focus on mean changes or variability
    - Ensure the quality of measurement
        - Asses sensitivity
        - Assess reliability
        - Assess validity

4. Analyze and interpret the evaluation results

5. Use information to modify training system, personnel selection, etc.

We learned important lessons at each step of the resource management evaluation process:

- An overall framework or theory about the type of performance measured must guide evaluation. Theory is particularly critical for specifying the levels for the expected effects, operationally defining resource management, and for specifying the other measures necessary to establish construct validity.

- The level of evaluation of resource management is often determined by the context. In commercial aviation, the most important levels of evaluation were the individual pilots and the flight crews.
- More controlled evaluation is desirable, but the selected evaluation design should be a workable compromise between the desire for control and the reality of the training and evaluation setting.
- A long-term multi-measure evaluation plan is necessary to detect delayed effects of training that may not be immediately apparent. Depending on the type of training the multiple measures may be over weeks, months, or years (as in the example study).
- Having a control (untrained) group helps avoid many confounds that would otherwise hamper single-group evaluations of training effects.
- The effects of resource management training should be examined in a broad range of possible changes. At a minimum, changes in knowledge, attitude, and behavior ought to be assessed.
- For measuring key effects like performance, multiple converging lines of evaluation evidence provide stronger support to the effects of resource management training. Creatively consider different ways the expected effects could be exhibited by individuals, teams, or organizational units.
- Repeated training in the evaluation process is necessary to maintain calibration of the raters for complex behavior domains such as resource management. Calibration must be checked for sensitivity, reliability, and validity wherever possible.
- Evaluation of resource management is an iterative process. That is, ongoing evaluation may cycle back from Step 5 to an earlier step in the process. In our research, results from year 1 and 2 LOE evaluations helped change the LOE evaluation format for more precise, comparative evaluations in year 3.

- Careful evaluation of resource management will result in a bonus of new knowledge about performance appraisal, the training program, and relevant individual and team processes. Our research uncovered new information about pilots, crews, and the organization.
- Careful choices must be made in each step of the evaluation process. Each choice involves trade-offs between the desire for the best possible evaluation of resource management and the constraints of time, personnel, and other critical resources.

Acknowledgments

This research was supported by the Office of the Chief Scientific and Technical Advisor for Human Factors (AAR-100) at the Federal Aviation Administration through grant 94-G-034 to George Mason University. We thank Dr. Eleana Edens at AAR-100 and Dr. Thomas Longridge at AFS-230 for their continuing support of this work.

References

Alliger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. Personnel Psychology, 42(2), 331-342.

Alliger, G. M., & Katzman, S. (1997). When training affects variability: Beyond the assessment of mean differences in training evaluation. In J. K. Ford & Associates (Eds.), Improving training effectuveness in work organizations. (pp. 223-246). Mahwah, NJ: Lawrence Earlbaum.

Alliger, G. M., Tannenbaum, S. I., Bennett, W., Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. Personnel Psychology, 50(2), 341-358.

Anderson, J. R. (1985). Cognitive psychology and its implications. New York: W. H. Freeman and Company.

Boehm-Davis, D. A., Holt, R. W., & Seamster, T. L. (this volume). Airline experiences with resource management programs. In E. Salas, C. Bowers & E. Edens (Ed.), Applying resource management in organizations: A guide for training professionals. Mahwah, NJ: Lawrence Erlbaum Aoccociates.

Boehm-Davis, D. A., Holt, R. W., & Hansberger, J. (1997). Pilot abilities and performance. In Procedings of the Ninth International Symposium on Aviation Psychology. Columbus, OH.

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), Personnel Selection in Organizations (pp. 71-98). San Francisco, CA: Jossey Bass.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56(2), 81-105.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Boston, MA: Houghton Mifflin.

Campbell, J. P. (1990). Modelling the performance prediction problem in industrial and organizational psychology. In M. Dunnette & L. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., pp. 687-732). Palo Alto, CA: Consulting Psychologists Press.

Cook, T. D., & Campbell, D. T. (1976). Quasi-experimentation: Design & analysis issues for field settings. Boston, MA: Houghton Mifflin.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology, 78(1), 98-104.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika , 16 , 297-334.

DeNisi, A. S., Cafferty T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: a model and research propositions. Organizational Behavior and Human Decision processes, 33, 360-396.

Dubois, C. L. Z., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximal performance criteria: Definitional issues, prediction, and white-black differences. Journal of Applied Psychology, 78, 205-211.

Goldstein, I. L. (1993). Training in organizations (3rd edition). Pacific Grove, CA: Brooks/Cole Publishing.

Goodman, P. S., Lerch, F. J., & Mukhopadhyay, T. (1994). Individual and organizational productivity: Linkages and processes. In D. H. Harris (Ed.), Organizational linkages: Understanding the productivity paradox (pp. 54-80). Washington DC: National Academy Press.

Guttentag, M., & Struening, E.L. (1975) Handbook of Evaluation Research (Volume 2) Beverly Hills, CA: Sage Publications

Hays, W. L. (1988). Statistics (4th edition). Chicago: Holt, Rinehart and Winston, Inc..

Helmreich, R. L., & Foushee, H. C. (1993). Why crew resource management? Empirical and theoretical bases of human factors training in aviation. In E. L. Weiner, B. G. Kanki, & R. L. Helmreich (Eds.), Cockpit resource management (pp. 3-45). San Francisco: Academic.

Holt, R. W., Johnson, P. J., & Goldsmith, T. E. (1997). Application of psychometrics to the calibration of air carrier evaluators. In Procedings of the Ninth International Symposium on Aviation Psychology. Columbus, OH.

Howell, D. C. (1997). Statistical methods for psychology (4th edition). Boston, MA: Duxbury Press.

Joint Committee on Standards for Education Evaluation. (1994). The Program Evaluation Standards (2nd edition). Sage, Thousand Oaks, CA.

Kirkpatrick, D. L. (1976). Evaluation of training. In R. L. Craig (Ed.), Training and development handbook: A guide to human resource development (2nd edition). New York: McGraw-Hill.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based and affective theories of learning to new methods of training evaluation. Journal of Applied Psychology [Monograph], 78(2), 311-328.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. American Psychologist, 41(11), 1183-1192.

Lauber, J. K. (1984). Resource management in the cockpit. Air Line Pilot, 53, 20-23.

Lawshe, C. H. (1975). A quantiative approach to content validity. Personnel Psychology, 28, 563-575.

National Transportation Safety Board. (1994). A review of flightcrew-involved, major accidents of U.S. air carriers, 1978 through 1990. (Safety Study NTSB / SS-94 / 01, Notation 6241): National Transportation Safety Board, Nunally, J.C. (1967) Psychometric Theory. New York, NY: McGraw-Hill.

Pedhazur, E. J., & Pedhazur Schmelkin, L. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Earlbaum.

Prince, C., Oser, R., Salas, E., & Woodruff, W. (1993). Increasing hits and reducing misses in CRM/LOS scenarios: Guidelines for simulator scenario development. International Journal of Aviation Psychology, 3(1), 69-82

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximal performance. Journal of Applied Psychology, 73(3), 482-486.

Steiner, I. D. (1972). Group process and productivity. New York: Academic Press.

Wiener, E. L., Kanki, B. G. , & Helmreich, R. L. (1993), Cockpit resource management. San Francisco: Academic.

Williams, D. M., Holt, R. W., & Boehm-Davis, D. A. (1997) Training for inter-rater reliability: baselines and benchmarks. In Procedings of the Ninth International Symposium on Aviation Psychology. Columbus, OH.